行政院國家科學委員會補助專題研究計畫成果報告

小樣本多變數下選取重要變數之研究

計畫類別：個別型計畫
計畫編號：NSC 97-2118-M-009-001-MY2
執行期間： 97 年 8 月 1 日至 99 年 7 月 31 日

執行機構及系所：國立交通大學統計學研所

計畫主持人： 洪慧念

計畫參與人員：吳侑峻 李博文 陳羽偉 林士傑 侯宏興

成果報告類型：完整報告

本計畫除繳交成果報告外，另須繳交以下出國心得報告：
●赴國外出差或研習心得報告
□赴大陸地區出差或研習心得報告
□出席國際學術會議心得報告
□國際合作研究計畫國外研究報告

處理方式：除列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

中 華 民 國 年 月 日

中文摘要

近十多年來由於基因晶片的發明產生了大量高密度的 cDNA 陣列資料。這些資料有著共同的特性就是樣本數不多但是基因數目很多。解決這類的問題，可以分成兩的步驟。首先是如何挑選重要的基因，接著是要如何的利用這些基因做分析。在本計畫中，我們對這兩方面做些有系統性的理論研究。在這些問題上，理論結果並不多。Fan等人近幾年提出關於如何選取恰當的基因數的理論依據，他們選取基因數目的準則是希望分類成功率愈高愈好。幾年前， Bickel 等人證明如果選取太多的基因，在分類上都不會有太好的結果。在理論結果中，共通的假設是當測量的基因數目愈多時，影響某特殊疾病的基因數目也成一定的方式迅速增多，且觀察的樣本數也以一定的方式增多。在本計畫中，我們討論當樣本數固定時，可測得的基因數目增加很快。倘若影響某疾病的基因數目也固定（或以非常慢的數度增加)，我們應該選取多少數目的基因以做資料分析最爲恰當。在本計畫中的另一個重點在於對 Tibshirasni 與 Tastie 於2007發表的文章做更深入的坦討與改進。在他們的研究中假設重要致病基因並不是對所有病人皆會有異常的表現，大約有（20%~100%) 的病人在此基因會有較正常人強烈的表現。對於他們的方法我們認爲還有不少可以討論與改進個空間。同時，我們也採用一些混和的模型對基因表現有異常的人數做出估計。

英文摘要

With advance technology in biology, high-throughput data such as microarry data are frequently seen in research work. Those data sets usually contains only a few samples but large number of variables. For analyzing this kind of data, fist we need to rank the importance of variables (genes), then we need to choose an importance subset of variables (genes) to analyze the microarray data (classification problem). In this two-year project, we will try to solve these two problems systematically and find some theoretical results. For these problems there are only few theoretical results. Recent years, some researchers find good theoretical results about find a good subset of important genes. Many years ago, Bickel showed that if we use too many genes to do classification problem, the Fisher discriminant performs poorly. All the theoretically results, under large sample, assume that when the number of variables (genes) goes to infinity, the number of sample in normal group and disease group are both go to infinity. Also the number of the important variables (genes) goes to infinity. In this project, we will discuss the situation when the number of sample size is fixed and the number variables (genes) goes to infinity. Also, we will assume that the number of important genes is fixed (or goes to infinity in a slow speed). Under above assumptions, we will try to find a good subset of genes to do our data analysis. Another purpose of this project is to extend the result by Tibshirasni and Tastie (2007). In their paper, they assume that only part of the people (20%~100%) in disease group has abnormal gene expression. We hope that we can extend their method and then find a better statistic to rank the importance of the variables (genes).

關鍵詞:小變數大樣本 基因選取 t-分配

報告內容

Contents

1 Introduction

The microarray data in biomedical research has been studied extensively in the past few years. Microarray is a technology to detect mRNA expression level. In general, detecting mRNA expression level can help identify genes that contribute to disease. That is, the goal of a microarray experiment is to identify those genes that are differentially expressed within different samples. Besides, the number of samples we observed is much less than the number of genes in a microarray experiment, thus generating a large-scale multiple hypothesis testing problem (Gentleman, Carey, Huber, Irizarry, and Dudoit, 2005; Efron, 2007).

A large-scale multiple hypothesis testing problem in a microarray experiment involves the simultaneous test of thousands, or even millions, of null hypotheses (Gentleman et al., 2005). Usually we use two-sample t-statistics t_i comparing expression levels under two different conditions for m genes. Then, the t_i's are transformed to z_i's such that, under normal assumption, z_i has a standard normal distribution (Efron, 2007). Efron (2007) displayed two histograms of z_i's from two microarray experiments and described the z_i's correlations can cause the fact that the distribution of the z_i's differs from N(0,1), called theoretical null distribution.

Since the earlier study did not focus on the reason of the histograms of z_i's differing from N(0,1) on multiple testing procedures. Hence, in this paper, we have two purposes: (a) to discuss the possible reasons for the distribution of the z_i's differing from N(0,1); (b) to simulate the data from the possible models and recommend the possible reasons in large-scale multiple hypothesis testing problem.

2 Literature Review

Multiple Hypothesis Testing in a Microarray Experiment

Suppose we have a microarray experiment which produces gene expression data on m genes for n mRNA samples. Then the gene expression levels may be summarized by a m × n matrix X =(x_{ij}), where x_{ij} denotes the expression measures of gene i and sample j. The rows i =1,...,m represent the prob sets and the columns j =1,...,n represent the different microarrays. The gene expression levels might be either absolute or relative to the expression levels of a suitably defined common reference sample.

In a microarray experiment, the number m is usual several thousands or even millions and the number n is usual anywhere between around eight and a few hundreds. In a typical experiment, the n samples would consist of n_1 treatment samples and n_2 control samples, for example, the treatment samples are patients with BRCA1 mutations and the control samples are patients with BRCA2 mutations in breast cancer study. The goal of a microarray experiment is to identify those genes that are differentially expressed in the different mutations of breast cancer. Therefore, suppose the single test is considered for each gene, the null hypothesis for testing that the gene i has the same expression distribution under two different conditions. For tests of means, the test statistic is the usual two-sample t-statistic, where the two-sample t-statistic depends on the standard t-test for Welch t-test. Thus, we have m null hypotheses to consider simultaneously, each with its own test statistic,

Null hypothesis : $H_1, H_2, \ldots, H_i, \ldots, H_m$

Test statistic : $t_1, t_2, \ldots, t_i, \ldots, t_m$.

Then, we transform t_i to a z_i such that, under normal assumption, z_i has a standard normal distribution and derive rejection regions (Gentleman et al., 2005). The adjusted p-value for null hypotheses is defined as the smallest type I error, α, FWER or FDR, at which one would reject H_i in the multiple hypothesis testing problem. Finally, we reject the null hypotheses if the adjusted p-value is smaller than α. That is to say, we reject the H_i, means that the gene i is differentially expressed under two different mutations of breast cancer. The procedure of the several tests with controlled in type I error is called a multiple testing procedure, abbreviated MTP.

It is noteworthy that Benjamini and Hochberg (1995) de□ned the FDR to be the expected proportion of true null hypotheses among the rejected hypotheses, FDR = E(V/R), where V denote the number of rejecting H_0 under H_0 is true and R denote the number of rejecting H_0 in all hypotheses. Besides, Efron et al. (2001) and Efron (2004) described that local false discovery rate, $fdr(z)=f_0(z)/f(z)$, is closely related to Benjamini and Hochberg's FDR criterion. The density $f_0(z)$ is null probability density function (e.g., theoretical, empirical, or permutation null hypothesis distribution) and the density $f(z)$ is probability density function derived from the empirical distribution of the z_i's. Moreover, Efron (2004) report that we can find out the genes which are differentially expressed by the local fdr. The details about local fdr are described in Efron (2004) and Efron et al. (2001).

The choice of null distribution (e.g., theoretical, empirical, or permutation null hypothesis distribution) is important to control the local fdr (Efron 2004, 2006, 2007; Gentleman et al., 2005). Di□erent choices may in□uence the conclusion on identifying which genes as differential or the same in the multiple hypothesis testing (Efron 2004, 2006, 2007; Gentleman et al., 2005). Efron (2004) reported that the appropriate choice of null distribution is the empirical null rather than the theoretical null or permutation null in some microarray experiments. Also, Efron (2006) suggested that the theoretical null or permutation null is inappropriate null in HIV study since the theoretical null or permutation null may make there is no differential genes on MTP (Efron, 2006). Hence, we need to select a suitable distribution in multiple hypothesis testing under different microarray experiments.

Microarray Experiments

For the microarray experiments, we consider the breast cancer study and the HIV study below.

The Breast Cancer Study

Hedenfalk, Duggen, Chen, et al. (2001) reported on a microarray experiment concerning the mutant genes of hereditary breast cancer. It is known that two different mutations, BRCA1 and BRCA2, lead to greatly increased breast cancer risk.

The experiment included 15 breast cancer patients, 7 from BRCA1 mutation patients and 8 from BRCA2. Each patient measured a microarray of expression levels for the same m = 3226 genes. Then, we have a $m \times n$ matrix $X = (x_{ij})$ for the breast cancer study, where $m = 3226$ rows denote genes and $n = 15$ columns denote microarrays. Each row of X (i.e., gene) yielded a two-sample t-statistic t_i comparing BRCA1 with BRCA2 patients, which was then transformed to a z_i.

$$z_i = \Phi^{-1}(G_0(t_i)), i = 1, 2, \ldots, m,$$

where Φ is the standard normal cumulative distribution function (c.d.f.), and G_0 is the c.d.f. of a standard Student's t distribution with 13 degrees of freedom. Hence, we get $m = 3226$ test statistic z_i's and the distribution of the z_i's are displayed in Figure.

The HIV Study

The human immunodeficiency virus (HIV) study, described by van't Wout et al. (2003), contained 8 samples, 4 from HIV-positive patients and 4 from HIV-negative controls. Each samples measured a microarray of expression levels for the same $m = 7680$ genes. Then, we have a $m \times n$ matrix $X = (x_{ij})$ for the HIV study, where $m = 7680$ rows denote genes and $n = 8$ columns denote microarrays. Each row of X (i.e., gene) yielded a two-sample t-statistic t_i comparing HIV-positive patients with HIV-negative controls,
which was then transformed to a z_i.

$$z_i = \Phi^{-1}(G_0(t_i)), i = 1, 2, \ldots, m,$$

where Φ is the standard normal c.d.f., and G_0 is the c.d.f. of a standard Student's t distribution with 6 degrees of freedom. Hence, we get $m = 7680$ test statistic z_i's and the distribution of the z_i's are displayed in Figure 1(b) (Efron, 2004, 2005, 2006, 2007; Gottardo et al., 2006).

The data from the breast cancer study and the HIV study were two-color cDNA microarrays and people make quality assessment and preprocessing (e.g. normalization) for the data before using them in multiple hypothesis testing (Dudoit et al., 2003; Gottardo et al., 2006; Gentleman et al., 2005).

Efron (2007) described that we usually presuppose most of the genes to be null in microarray experiments, the goal being to identify some signi□cant nonnull genes. Therefore, we expect z_i to have closely a standard normal distribution for null genes (Efron, 2007). In other words, under null hypothesis, z_i should have a standard normal distribution if gene i has the same expression distribution for BRCA1 and BRCA2 patients or for HIV-positive patients and HIV-negative controls. Efron (2007) reported that heavy curves indicate N(0,1) theoretical null densities and light

curves indicate empirical null densities □t to central z-values in Figure, as done by Efron (2004). However, the histograms of z-values in Figure, where the distribution of the z_i's from breast cancer is wider than N(0,1) and from HIV study is narrower than N(0,1) (Efron, 2006, 2007). Efron (2007) pointed out that the correlations in multiple hypothesis testing can make the observed all z_i's behave as N(0, σ^2), where σ is obviously different than 1. Next section, we will discuss the correlation and other reasons for this phenomenon.

3 The Empirical Distribution of the z_i's

In this section, we discuss the possible reasons which caused the distribution of the z_i's that obviously di□ers from the N(0,1) in microarray experiments. First, Efron (2007) indicated that there were some gene correlations in the breast cancer data and in the HIV data. Besides, the disease is caused by abnormal genes and there are essential correlations between genes in biology. Hence we may say that there are gene correlation structures in the breast cancer data and the HIV data.

Secondly, Hedenfalk et al. (2001) pointed out that these patients with primary breast cancer and who had a family history of breast or ovarian cancer or both were asked to provide a blood sample for BRCA1 and BRCA2 mutations in the genetic breast cancer. If some of the patients are come from the same family, some of their gene may correlate. Hence the patients may correlate with the relationship of relatives.

Furthermore, Efron (2004) indicated that the □rst four and the last four microarrays in the BRCA2 patients were mutually correlated. Moreover, since the HIV is a rare disease, the HIV patients usually have the same features, for example, the patients are homosexuality, drug addicts and infected with mother. According to the above, we may safely say that there are the correlation structures among patients (i.e. microarrays).

Finally, if the data (x_{ij}) are independent and identically distributed (i.i.d.) random variables from normal distribution, we may apply the two-sample t-statistic in multiple hypothesis testing. In other words, if the data (x_{ij}) are independent and identically distributed (i.i.d.) random variables from other distributions, the two-sample t-statistic may not have the t-distribution.

Hence, as mentioned above, we may consider the three possible reasons under the following items :

(1) correlation between genes. (2) correlation among microarrays. (3) various distribution assumptions. In the next section, we discuss further the models of these possible reasons. Besides, we apply these models for simulating data and then compare the results of the simulation.

4 The Models and Simulation Study

For generating dependent data, we consider two kinds of time series models: the autoregressive model (AR) and the moving average model (MA). We introduce the AR model and the MA model.

Definition 1 An autoregressive model of order p, abbreviated AR(p), is defined to be

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t,$$

where X_t is stationary, $\phi_1, \phi_2, \ldots, \phi_p$ ($\phi_p = 0$) are constants, and Z_t is a Gaussian white noise series with mean 0 and variance σ^2.

Definition 2 A moving average model of order q, abbreviated MA(q), is defined to be

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q},$$

where there are q lags in the moving average, $\theta_1, \theta_2, \ldots, \theta_q$ ($\theta_q = 0$) are constants, and Z_t is a Gaussian white noise series with mean 0 and variance σ^2.

Suppose a microarray experiment includes n ($n = n_1 + n_2$) patients, n_1 from group 1 and n_2 from group 2. Each patient measures a microarray of expression levels for the same m genes. We want to identify those genes that are differentially expressed under the two group. Let $X = (x_{ij})$ represent gene expression and be a $m \times n$ matrix, where $i = 1, \ldots, m$ denotes genes and $j = 1, \ldots, n$ ($n = n_1 + n_2$) denotes microarrays.

In the simulation study, we choose $m = 100000$ genes and $n = 14$ ($n_1 = n_2 = 7$) micrarrays. Then we apply the data on the multiple testing procedures. Therefore, we get $m = 100000$ z_i's.

Models of correlation between genes

In the following models, we consider that there is some correlation between genes, but there is no dependence between microarrays.

Model 1

For model 1, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1} \square$ i.i.d. $N(0, \sigma^2)$ $x_{1j}, x_{2j}, \ldots, x_{mj} \square$ AR(p), $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in} \square$ i.i.d. $N(0, \sigma^2)$,

Model 2

For model 2, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1} \square$ i.i.d. $N(0, \sigma^2)$ $x_{1j}, x_{2j}, \ldots, x_{mj} \square$ MA(q), $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in} \square$ i.i.d. $N(0, \sigma^2)$,

Model 3

For model 3, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1} \square$ i.i.d. $N(0, \sigma^2)$

$x_{in_1+1}, x_{in_1+2}, \ldots, x_{in} \square$ i.i.d. $N(0, \sigma)$,

$cor(x_{kj}, x_{lj}) = c$, $k = 1, \ldots, m$, $j = 1, \ldots, m$, $k = 1$,

Model 4

We consider $x_{i1}, x_{i2}, \ldots, x_{in_1} \square$ AR(p)

$x_{1j}, x_{2j}, \ldots, x_{mj} \square$ i.i.d. $N(0, \sigma^2)$, $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in} \square$ AR(p),

Model 5

$x_{i1}, x_{i2}, \ldots, x_{in_1} \square$ MA(q) $x_{1j}, x_{2j}, \ldots, x_{mj} \square$ i.i.d. $N(0, \sigma^2)$, $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in} \square$ MA(q),

Model 6

For model 6, we consider

$x_{1j}, x_{2j}, \ldots, x_{mj} \square$ i.i.d. $N(0, \sigma^2)$, $cor(x_{ik}, x_{il}) = c$, $k = 1, \ldots, n_1, j = 1, \ldots, n_1, k = 1$
$cor(x_{ik}, x_{il}) = c$, $k = n_1 + 1, \ldots, n$, $j = n_1 + 1, \ldots, n$, $k = 1$,

Model 7

For model 7, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1}$ □ i.i.d. Gamma(α = shape, λ = rate), $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in}$ □ i.i.d. Gamma(α = shape, λ = rate),

Model 8

For model 8, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1}$ □ i.i.d. Cauchy(α = location, λ = scale)

$x_{in_1+1}, x_{in_1+2}, \ldots, x_{in}$ □ i.i.d. Cauchy(α = location, λ = scale),

Model 9

For model 9, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1}$ □ i.i.d. W eibull(λ = shape, α = scale, β = location)

$x_{in_1+1}, x_{in_1+2}, \ldots, x_{in}$ □ i.i.d. W eibull(λ = shape, α = scale, β = location),

Model 10

For model 10, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1}$ □ i.i.d. Exp(λ = rate) $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in}$ □ i.i.d. Exp(λ = rate),

Model 11

For model 11, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1}$ □ i.i.d. t(n = degrees of freedom) $x_{in_1+1}, x_{in_1+2}, \ldots, x_{in}$ □ i.i.d. t(n = degrees of freedom),

Model 12

For model 12, we consider

$x_{i1}, x_{i2}, \ldots, x_{in_1}$ □ i.i.d. F (v_1, v_2)(v_1, v_2 = degrees of freedom)

$x_{in_1+1}, x_{in_1+2}, \ldots, x_{in}$ □ i.i.d. F (v_1, v_2)(v_1, v_2 = degrees of freedom),

5 Real Data

The data is a microarray experiment about breast cancer, which provided by Department of Interdisciplinary Oncology Mo□tt Cancer Center and Research Institute, University of South Florida. The experiment included 185 samples, 143 from the normal group and 42 from the patients. Each samples measured a microarray of expression levels for the same m = 54675 genes. Then we apply the data on the multiple testing procedures and therefore we get m = 54675 z_i's. The histogram of the observed z_i's plot is in the Figure 11. In Figure 11, heavy blue line indicates the theoretical null distribution. We can see that the empirical distribution of the z_i's is more wide than the N(0,1). Hence, we guess that the data may have correlation among microarrays. Also, if the genes are null, these z_i's should have a standard normal distribution under normal assumption. In order to solve the problem, we may try some improved method. For example, permutation methods can be used to avoid the assumption of $z_i|H_i$ □ N(0,1) and possibly make the permutation-improved theoretical null will more closely match the empirical null (Efron et al. 2001; Dudoit et al. 2003; Efron 2004; Efron 2007). Moreover, Efron (2007) referred to the random permutation of the microarrays can eliminate the group di□erences and preserve the correlation structure of the genes. Hence we apply permutation methods to the breast cancer data.

Let X represent the 54675 × 185 matrix X =(x_{ij}) of the breast cancer data. Each row of X (i.e., each gene) yields a two-sample t-statistic t_i comparing 143 from the normal group and 42 from the patients, which is then transformed to a z_i by $z_i = \Phi^{-1}(G_0(t_i))$ and we get 54675 z_i's. Then, we recalculate the 54675 z_i's by randomly permuting the columns of X. Namely, we recalculate the 54675 z_i's by randomly dividing the 185 samples into groups of 143 and 42. This process is independently repeated 100 times, generating a total of 100 × 54675 permutation z_i's. This testing is called permutation testing. Since permutation test is model-free, we can say that permutation test is more robust than t-test. The empirical distribution of the 100 × 54675 z_i's (i.e., permutation null) plot is in the Figures, heavy red line indicates the distribution of the 100 × 54675 z_i's (i.e., permutation null). We can see that the empirical distribution of the z_i's is more wide than the permutation null distribution, but the permutation null is more closely match the histogram of the observed z_i's than the N(0,1).

However, permutation methods are a way of avoiding the normal assumption (Dudoit et al., 2003; Efron, 2001, 2004, 2006), but they do not solve the problem of selecting a suitable null hypothesis (Efron, 2004). The choice of a suitable null hypothesis can see Efron (2004, 2006, 2007).

6 Conclusions and Future Research

In this study, we focused on the reasons of empirical distribution of the z_i's differed from N(0,1) in large-scale multiple hypothesis testing. We proposed the three possible reasons. The first reason was the correlation between genes. The secondly reason was the correlation among microarrays. The third reason was the various distribution assumptions. Moreover, we provided twelve models from three different reasons and simulated the data by the models.

By observing the simulated data from models of correlation among microarrays, we could see that the empirical distribution of the z_i's may differs from N(0,1) as the correlation getting larger. Also, we see that there is a significant difference between the empirical distribution of the z_i's and the N(0,1) by observing the simulated data from models of various distribution assumptions. Hence, by the simulation results we conclude that the correlation between genes could not affect the empirical distribution of the z_i's and that the correlation among microarrays and various distribution assumption are the main reasons.

This study only proposed three possible reasons in large-scale multiple hypothesis testing. It might be worth to discuss further possible reasons that may make the distribution of the z_i's differing from N(0,1) and provide appropriate models for the other possible reasons.

Also, this study used the AR and MA model with different coefficients and order to generate the correlation data between genes and among microarrays. Another direction for future research is to use an autoregressive moving average (ARMA) model or other correlation model for the proposed reasons. In addition, this study provided six different distribution models for the various distribution assumptions. It might be assume other distribution models to investigate further in future research.

考文獻

1. Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* 19 563-570.

2. Buhlmann, P. and Bin, Y. (2004). Discussion of boosting papers. Ann. Statist. 3296–101.

3. Bai, Z. and Saranadasa, H. (1996). Effect of high dimension : by an example of a two sample problem. Statistica Sinica 6, 311-329.

4. Bair, E., Hastie, T., Debashis, P., and Tibshirani, R. (2007) Prediction by supervised principal components. The Annals of Statistics

5. Buhlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classication. Journal of the American Statistical Association 98, 324-339.

6. Bickel, P. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, naive Bayes, and some alternatives where there are many more variables than observations. Bernoulli 10 989–1010.

7. Boulesteix, A. (2004). PLS Dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology* 3 1-33.

8. Breiman, L. (2001). Statistical modeling: The two cultures (with discussion). Statist.Sci. 16 199–231.
9. Breiman, L. (2004). Population theory for boosting ensembles. Ann. Statist. 32 1–11.8
10. Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. Bioinformatics, 19,1252-1258.
11. Cao, H.Y. (2007). Moderate deviations for two sample t-statistics. Probability and Statistics.
12. Chen, S., Donoho, D. and Saunders, M. (2001). Atomic decomposition by basis pursuit. SIAM Rev. 43 129–159.
13. Dettling, M. and Buhlmann, P. (2003). Boosting for tumor classification with gene expression data. Bioinformatics 19 No. 9, 1061-1069.
14. Donoho, D. (2004). For most large underdetermined systems of linear equations of minimal l1-norm solution is also the sparsest solution. 2004-9, Dept. Statistics, Stanford Univ.
15. Donoho, D. (2004). For most large undetermined systems of equations, the minimall1-norm near-solution approximates the sparsest near-solution.
16. Dudoit, S., Fridlyard, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97 77-87.
17. Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004). Least angle regression (with discussion). Ann. Statist. 32 407–499.
18. F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. Mathematical Biosciences176 123{144.
19. Fan, J and Ren, Y. (2006). Statistical analysis of DNA microarray data. Clinical Cancer Research 12 4469-4473.
20. Fan, J. (1996). Test of significance based on wavelet thresh holding and Neyman's truncation. Journal of the American Statistical Association 91 674-688.
21. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 1348–1360.
22. Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proceedings of the International Congress ofMathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III,595-622.
23. Fan, J. and Lv, J. (2007). Sure independence screening for ultra-high dimensional
24. Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32 928–961.
25. Fan, J., Hall, P. and Yao, Q. (2006). To how many simultaneous hypothesis tests can normal, student's t or Bootstrap calibration be applied
26. Friedman, J. H. (1989). Regularized discriminant analysis. Journal of the American Statistical Association 84, 165-175.
27. Friedman, J., Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004). Discussion of boosting papers. Ann. Statist. 32 102–107.
28. Ghosh, D. (2002). Singular value decomposition regression modeling for classification of tumors from microarray experiments. Proceedings of the Pacific Symposium on Biocomputing, 11462-11467.
29. Greenshtein, E. (2005). Prediction, model selection and random dimension penal-ties. Sankhy a 67 46–73.
30. Greenshtein, E. (2006). Best subset selection, persistence in high dimensional statistical learning and

optimization under l1 constraint. Ann. Statist., 2367-2386
31. Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of over parametrization. Bernoulli 10 971-988.
32. Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer, New York.
33. Zou, H., Hastie, T., and Tibshirani. R. (2004). Sparse principal component analysis.
34. Zou, H., Hastie, T., and Tibshirani. R. (2007). Outlier sums for differential gene expression analysis, Biostatistics
35. Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data. Bioinformatics 19 2072-2978.
36. Huber, P. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo Ann. Statistics 1 799–821.
37. Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. Ann. Statistics. 28 681–712.
38. Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. IEEE Trans. Information Theory 42 2118–2132.
39. Lin, Z. and Lu, C. (1996). Limit Theory for Mixing Dependent Random Variables. Kluwer Academic Publishers.
40. Lugosi, G. and Vayatis, N. (2004). On the Bayes risk consistency of regularized boosting methods. Ann. Statistics 32 30–55.
41. Meinshausen, N. (2007). Relaxed Lasso. Computational Statistics and Data Analysis
42. Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. Ann. Statistics 34 1436–1462.
43. Nemirovski, A. and Yudin, D. (1983). Problem Complexity and Method Efficiency in Optimization. Wiley, New York.
44. Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18 39-50.
45. Nguyen, D. V., Arpat, A. B., Wang, N. and Carroll, R. J. (2002). DNA microarray experiments: Biological and technological aspects. Biometrics 58 701-717.
46. Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. *Seminar on Functional Analysis, 1980--1981*, *École Polytechnic*, *Palaiseau*. Exp. no. V, 13 pp.
47. Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p2/n is large. I. Consistency. Ann. Statistics 12 1298-1309.
48. Shao, Q. M. (2005). Self-normalized Limit Theorems in Probability and Statistics.
49. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal Statistics Soc. Ser. B 58 267-288.
50. Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. 99 6567-6572.
51. van der Vaart, A. W. and Wellner, J. A. (1996). Weak Convergence and Empirical Processes. Springer-Verlag, New York.
52. Vapnik, N. V. (1998). Statistical Learning Theory. Wiley, New York.
53. West, M., Blanchette, C., Fressman, H., Huang, E., Ishida, S., Spang, R.,Zuan, H., Marks, J. R. and Nevins, J.

R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. Proc. Natl. Acad. Sci. 98 11462-11467.
54. Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of M-estimators for the linear model, Ann. Statistics